SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 29 (2011) 188 – 197

Procedia
Social and Behavioral Sciences

International Conference on Education and Educational Psychology (ICEEPSY 2011)

The use of item analysis for the improvement of objective examinations

Anna Siri, Michela Freddano*

** University of Genoa, Corso A. Podestà 1, Genoa 16128, Italy*

Abstract

In the standardized and objective evaluation of student performances, the item analysis is a process in which both students' answers and test questions are examined in order to assess the quality and quantity of the items and the test as a whole. All students from some classrooms of primary and middle school were selected to evaluate their performances by testing. On the basis of the analysis results the tests have been re-designed. The results emphasized that item analysis provides valuable information to the teachers to further item modification and future test development and offers educational tools to assist them.

Keywords:

Item analysis; standardized test; student assessment; teacher made test; teacher training; measures of central tendency.

Selection and/or peer-review under responsibility of Dr Zafer Bekirogullari.
Keywords: Type your keywords here, separated by semicolons ;

1. Introduction

Evaluation is an essential dimension of education and plays an important role in giving feedbacks to stakeholders. Students' assessment and evaluation are an integral part of the teaching and learning process (Popham, 2002; Trice, 2000).

According with Xu and Liu (2009) the teachers' knowledge in assessment and evaluation is not a static process but rather a complex, dynamic, and ongoing activity. These Authors also argue that this type of knowledge develops along a temporal continuum; teachers usually use their past assessment experience to plan their current assessment practices.

Among the different types of students' learning achievements and progress the multiple choice questions are globally the most utilized (Swanson et al., 2006; 2005).

Most classroom assessment involves tests that teachers have constructed themselves (Carroll & Moody, 2006; Boothroyd et al., 1992).

* Anna Siri. Tel.: +39 010 3537235; fax: +39 010 3533428.
E-mail address: anna.siri@unige.it

Selection and/or peer-review under responsibility of Dr Zafer Bekirogullari.
doi:10.1016/j.sbspro.2011.11.224

Further, teachers place more importance on their own assessment than on tests designed by others or on other data sources to evaluate/(or) measure grades and student progress (Boothroyd, et al., 1992; Williams, 1991).
Generally teachers receive little or no assessment training and support (Herman et al., 1984). In addition this type of preparation often don't focus on strategies to construct test or item-writing rules but only on large-scale test administration and standardized test score interpretation (Stiggins, 1991). Most teachers believe that they need strong measurement of skills (Wise et al., 1991).
In Italy the culture of evaluation among schools and self-evaluation within schools are gradually developing with the school autonomy system.
This article is a part of a large project aimed to improve the expertise of teachers of some primary and middle schools of Genoa to systematically use standardized and objective evaluation of student achievement.
The experience was conducted by the PhD students of the doctoral course in Evaluation of Educational Processes and Systems of the University of Genoa.
The specific aim of this study is to investigate the effect of the analysis of multiple choice questions designed by the teachers on the quality of the tests.

2. Theoretical Background

In the standardized and objective evaluation of student performances, the item analysis is a process in which both students' answers and test questions are examined in order to assess the quality and quantity of the items and the test as a whole. The current empirical research literature has worked hard to promote valid interpretation of the characteristics, validity and reliability of quality assessment practices (Sireci et al., 2006).
The crucial thing of preparing multiple choice tests is to construct good questions. This requires first of all understanding of the objectives of assessment, having good skills in writing the items and an excellent knowledge of the content. There are some guidelines supported by experimental or quasi-experimental designs, but these are usually not adhered to. So the results are the preparation and administration of faulty tests (Walsh, 2008; Haladyna, 2004).
Item analysis is an examination of a test after its administration (Remmers et al., 1967). The quality of a test depends upon each items of a test (Shrama, 2000; Freeman, 1962).
Item analysis allows us to observe the item characteristics, and to improve the quality of the test (Gronlund, 1993). Item revision (Lange, 1967) allows to identify items too difficult or too easy, items not able to differentiate between students who have learned the content and those who have not, or questions that have distracters not plausible. So teachers can remove them from the pool of items or change the items or modify instruction to correct a confusing misunderstanding about the content or adjust the way they teach.
Improving the skills in the test through item analysis can save time and energy on the part of teachers and test designers.
Two approaches are widely used for item analysis:

- the Classical Test Theory that utilizes two main statistics: the item facility index **(**the percentage of students that correctly answered the item) and the Discrimination index (the point-bacterial relationship between students' performance on individual item and total test score).
- the Item Response Theory (IRT) that describes both item statistics and student's ability with the assumption of correlation between the score on a single item and overall test performance. The IRT assumes that there is a correlation between the score gained by a candidate for one (measurable) item/test and their overall ability on the latent trait which underlies test performance (that we want to discover).

The analysis of student achievement and the item analysis were useful to study the validity and reliability of the tests before and after their administration. Validity concerns the relationship between the indicators (the items) and the indicated variable (the skill on reading literacy). A test, a question or an item is reliable when it really measures what the researcher/evaluator (teacher) want to notice. For further information please see Palumbo et al. (2006).

3. The study

Aim

The specific aim of this study is to investigate the effect of the analysis of multiple choice questions designed by the teachers on the quality of the tests.

Design

Under a special agreement, some schools formed two networks, actually existent, one on the East, and the other on the West of Genoa. Seventy-four teachers participated to the training with the aim to realize an evaluative tool to

assess students' learning performances and to improve the vertical curriculum. More specifically the training course allowed teachers have some basic elements to carry out student learning evaluation using valid objectives test items. This activities started from previous experiences of assessment tested by the training participants, as the figure below shows.

Figure n. 1 - The activities of the teacher training

Levels	Activity	Aims	Recipients
Macro level: theoretical framework.	General meetings, workshops.	Introducing the potential users into the scholastic evaluation project and providing theoretical common bases.	Voluntary teachers or teachers selected by the principals of the school of the Province of Genoa.
Meso Level: evaluative language.	Background analysis and analysis of needs to organize the experimental school activities.	Sharing common evaluative language.	Teachers and principals from the school networks.
Micro Level: specific Practices.	Experimental school activities.	Improving and putting into practice evaluative knowledge and skills.	Teacher teams within the schools of the networks.

The analysis of needs identified the following educational goals:
- knowing how to build structured tests of reading literacy starting from eventual previously constructed test;
- being able to process, analyze and interpret the results;
- being able to disseminate the results to different stakeholders, particularly the teachers' board.

Particularly the focus of the training was on the acquisition of skills to evaluate students' performances on reading literacy. As a matter of fact, reading literacy is transversal and related with all disciplines.
The methodology was centered on the collaborative and cooperative work between experts, trainers and teachers. Five seminars had done by experts to construct and assess some reading tests. Teachers retraced the entire process from the construction to the validation of evaluation tests, with the aim of acquiring knowledge and obtaining useful tools to reuse in educational practice.
A pre-test activity allowed the evaluation of tests' validity and reliability from the quantitative and qualitative points of view. Quantitatively teachers evaluated tests' validity and reliability to assess student performances analyzing the mean, the median, the mode and the standard deviation. Moreover teachers did an item analysis computing the following statistical indicators: facility index, selectivity index and distracter index. Qualitatively teachers evaluate the level of both the clarity and understanding of the tests.
On the basis of the pre-test activity, teachers prepared the final student tests on reading literacy. The experimental activity involved 708 students. The test administration was guided by methodological notes and was carried out in different ways and times in both networks. This activity encouraged the teachers to improve and provided further inspiration for future exchange activities.

Figure n. 2 - The participant and the type of tests engaged

<table>
<tr><th colspan="2">East Network (with pre-test)</th><th colspan="2">West network (without pre-test)</th></tr>
<tr><td>92 students of third grade (8 years old)</td><td rowspan="4">Narrative text
Informative text
Grammar exercises</td><td>19 students of kindergarten (5 years old)</td><td>Test the comprehensive processes</td></tr>
<tr><td>108 students of fourth grade (9 years old)</td><td>90 students of second grade (7 years old)</td><td>Narrative text</td></tr>
<tr><td rowspan="2">8 students of fifth grade (10 years old)</td><td>150 students of fifth grade (10 years old)</td><td rowspan="2">Narrative text
Informative text
Grammar exercises</td></tr>
<tr><td>161 students of seventh grade (12 years old)</td></tr>
</table>

Data elaboration was facilitated by an "ad hoc" tool that was built for each test and useful for performance and item analyses. This tool was immediate and easy to use and of great help for the analysis because it speeded up the reading of data. To facilitate teacher learning we also provided a bibliography on reading literacy and some evaluation guidelines on how to realize an evaluation report and how to communicate results to different stakeholders (students, professors, parents, public administration, experts).

The evaluation of the training effectiveness allowed the reflection on the activity planning and the design of new actions. Some teachers indicated appreciation for the opportunity to openly voice their opinions and to reflect on school evaluation. The teachers expressed hopes of increasing their knowledge about the school evaluation tools and about the acquisition of a common evaluative language. Moreover some teachers expressed the need to find how to achieve students to the tests.

Participants

In this study 108 students, both male and female students from the 4th primary schools level in the network, constituted the population of the study.

Method

The research is driven by the theory that good and bad items behave differently. Classical approach was adopted due to its simplicity.

The topic was reading literature.

Teachers (supported by the experts) computed the facility index and the selectivity index to analyze the difficulty and the power of discrimination of the questions/items, and the distracter index to analyze, within each question, the pull of the distracters (incorrect responses) instead the correct response. When the major part of the test is composed by too easy or too difficult items, there is no variance between the student results. Moreover the analysis of the distracters allowed to realize reliable multiple choice questions with a balanced level of attractiveness within items (a correct response, a great distracter and two weak incorrect responses).

For each test teachers modified or deleted questions with the evidence to be not sufficiently valid or reliable. After pre-test the tests were administered to the student population; the analysis of student achievement assessed definitively students' skills in reading literacy; particularly the analysis of the measures of central tendency and the item analysis allowed to monitor the effective capacity of the test to measure student achievement.

The following sub-paragraph deepens the use of item analysis by a case study.

Practical application of the analysis of item analysis: the case of testing 4th grade students of the East Network of the Genoese Schools.

The test based on the informative text "*Merci e persone viaggiano*" was administered to 108 students of the 4th grade of the schools of the Genoese East Network (of 9 years old). Four students with serious cognitive disabilities did the test but their performance were excluded from the final analysis. It consisted of a text of 423 words and provided16 multiple choice questions (4 responses). The following figure shows what skills on reading literature from the Theoretical Framework (2011) by INVALSI (Italian National Institute for the Educational Evaluation of Instruction and Training) were assessed by the test.

The map of the questions (Fig. n. 1) was realized by Prof. Gabriella Ravizza, the expert on realizing student tests on reading literacy who followed teachers during this project. Particularly, a test that is based on an informative text provides more questions of the type 1, 2 and 4 than that of the other types. Moreover a test for student of the 4th grade (of 9 years old) doesn't provide questions about the interpretation of the text (type n. 6) or the evaluation and reflection on the content of the text (type n. 7).

Figure n. 3 - The map of the questions

Type	Labels	***Tasks***	Questions
1	Recognizing and understanding the literal and figurative meanings of words and sentences; recognizing associations between words	Recognizing the meaning of a word in the context	3
		Recognizing the literal meaning of a word	5
		Recognizing the meaning of morphological changes	9
2	Identifying the explicit information in the text	Locating specific information (literal, synonymous or paraphrasic) explicitly in the text	1, 4, 15

3	Doing direct inference, obtaining information from one or several implicit information from the text and/or from the personal encyclopedia	Inferring and explaining the cause of an event or action	8
		Inferring the characters of an aspect	14
4	Identifying relationships of both text cohesion (logical organization within and beyond the sentence) and coherence	Identifying the reference (pronoun) of an anaphora	7, 13
		Recognizing the significance and function of explicit connective	10,12
5a	Reconstructing the meaning of a part of the text, integrating more information and concepts, also doing complex inferences	Integrating or linking information from the text and/or the personal encyclopedia	6, 11
5b	Reconstructing the global meaning of the text, integrating more information and concepts, also formulating complex inferences	Identifying the main topic of the text	16
6	Developing an interpretation of the text, starting from its content and/or form, going beyond the literal reading comprehension		--
7	Evaluating the content and/or the form of the text by using personal knowledge and experiences (reflecting on: the plausibility of the information; the validity of the argumentation; the communicative effectiveness of the text; etc.)		--

4. Results and discussion

The item analysis

Teachers computed the facility index and the selectivity index to analyze the items. Resuming the facility index is computed by subdividing the percentage of the correct answers to a question per 100. The value of the facility index ranges between 0 and +1. The facility of a question is acceptable if the value of the facility index ranges between .26 and .75. While the item is very easy if the value of the facility index is major than .75 or the item is very difficult if the value of the facility index is minor than .26.

The value of the selectivity index ranges between 0 and +1. An item is selective when the value of the selectivity index ranges between .30 and .60; while the item is too selective when the value of the selectivity index is major than .60, or not selective when the value of the selectivity index is minor than .30.

Table n. 1 - Percentage of correct and incorrect answers and no response for each item of the test

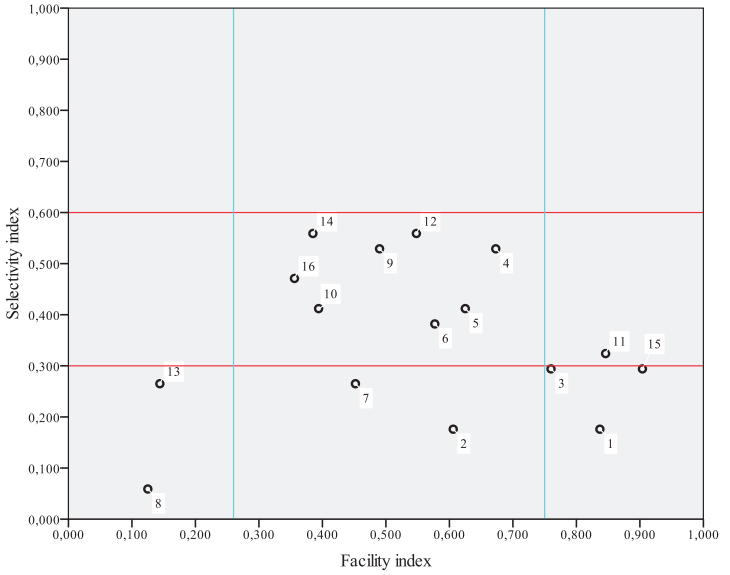
Question	Correct answers	Incorrect answers	No response	% correct	% incorrect	% No response	Facility index		Selectivity index	
1	87	16	1	83,65	15,39	0,96	,837	EE	,176	NS
2	63	40	1	60,58	38,46	0,96	,606	E	,176	NS
3	79	24	1	75,96	23,08	0,96	,760	EE	,294	NS
4	70	33	1	67,31	31,73	0,96	,673	E	,529	S
5	65	37	2	62,5	35,58	1,92	,625	E	,412	S
6	60	42	2	57,69	40,39	1,92	,577	E	,382	S
7	47	55	2	45,19	52,89	1,92	,452	D	,265	NS
8	13	87	4	12,5	83,65	3,85	,125	DD	,059	NS
9	51	52	1	49,04	50	0,96	,490	D	,529	S
10	41	58	5	39,42	55,77	4,81	,394	D	,412	S

11	88	14	2	84,62	13,46	1,92	,846	EE	,324	S
12	57	42	5	54,81	40,38	4,81	,548	E	,559	S
13	15	84	5	14,42	80,77	4,81	,144	DD	,265	NS
14	40	60	4	38,46	57,69	3,85	,385	D	,559	S
15	94	8	2	90,38	7,7	1,92	,904	EE	,294	NS
16	37	65	2	35,58	62,5	1,92	,356	D	,471	S

Legend: in the column "% correct" questions that obtain less than one half of correct responses are underlined. In the column "% no response" questions that obtain more than 2,5% responses are underlined. In the column "Facility index" the label E= easy; EE = very easy; D = difficult; DD = very difficult. In the column "Selectivity index" the label S = selective and the label NS = no selective.

Matching the index of both facility and selectivity on the Cartesian plan provided the graphical position of the item in front of its facility and selectivity. The graphical representation immediately allowed the identification of the appropriate (valid and reliable) questions at the center of the plan, where the values of both the indexes were on the average. Moreover the representation was useful to notice immediately the questions too/not easy or too/not selective that we called "outside". In our case the outside items n. 8 and 13 were too difficult and not selective, while the items n. 1, 11 and 15 were easy and not selective. Particularly the items n. 1 and 15 were easy because they were about the identification of explicit information in the text (code n. 2, fig. 1).

Graph n. 1 - Matching the facility index and the selectivity index



The analysis of the measures of central tendency

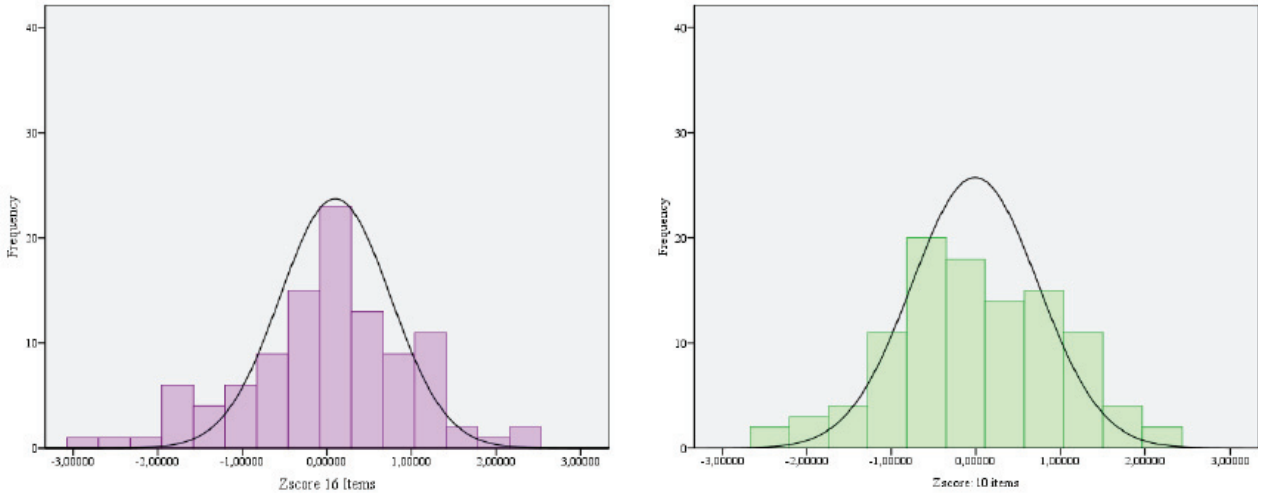
The final score ranged between 0 and 16 attributing one point to each correct answer. No student achieved the highest score (16). On average students correctly answered to more than one half of the test (mean of 8,72 and standard deviation of 2,675). The mode and the median were higher than the average overlapping at 9. The students' score distribution was asymmetrically negative (the high scores were more than the low scores). The analysis of the deciles showed that the test was difficult: the first decile of the student distribution correctly answered to a maximum of 5 questions (the 31,25% of the test); the second decile correctly answered to a maximum of 7 questions (the 43,75% of the test) and only from the third decile, students correctly responded at least to one half of the test.
The computation of the measures of central tendency of the tests with and without the items that are outside the central box of the Graph 1(the questions n. 1, 2, 8, 11, 13 and 15) was useful to compare the students' score

distribution. We computed the standardized values (Z scores) assuming the mean equal to 0 and standard deviation equal to 1. The table n. 2 shows the comparison of the measures of the test composed by 16 questions with that of 10 questions (without questions n. 1, 2, 8, 11, 13 and 15): in the second case the distribution was more normal (please see the values of the skewness) and less concentrate (please see the values of the kurtosis) than in the first case. In the second case, the measures of central tendency were minor than that in the first case and, particularly, in the second case the mean was major than the median and the mode: the students' score distribution was positively asymmetric (more low scores) in the second case than in the first case, where the distribution was negatively asymmetric (more higher scores). We could assume that the test was difficult because more than the half of the students did wrong the valid items.

Table n. 2 - The comparison by using normalized values

		Values (16 items)	Zscore 16 Items	Zscore 10 Items
Mean		8,72	,0	,0
Median		9	,104	-,121
Mode		9	,104	-,585
Std. Deviation		2,675	1	1
Skewness		-,331	,331	-,094
Kurtosis		,314	,314	-,288
Minimum		1	-2,887	-2,442
Maximum		15	2,347	2,201
Percentiles	10	5	-1,391	-1,049
	20	7	-,643	-,585
	30	8	-,270	-,585
	40	8	-,270	-,121
	50	9	,104	-,121
	60	9	,104	,344
	70	10	,478	,808
	80	11	,852	,808
	90	12	1,226	1,272

Graph n. 2 - Z scores' distribution of the integral test (16 items; in violet) and the 10 valid items (in green)



The distracter analysis to re-design questions

The distracter analysis allowed the study of the outside items (outside the central box, graph 1) analyzing the way in which students answered. The distracter index is computed subdividing the percentage of responses of a question per 100. The value of the distracter index ranges between 0 and +1. The next table shows the level of distraction of each question of the test.

Table n. 3 - The distracter analysis

Question	A	B	C	D	No response
1	,048	,837	,077	,029	,010
2	,337	,019	,029	,606	,010
3	,760	,058	,029	,144	,010
4	,144	,673	,029	,144	,010
5	,173	,173	,625	,010	,019
6	,096	,269	,577	,039	,019
7	,260	,452	,212	,058	,019
8	,154	,067	,615	,125	,039
9	,087	,490	,356	,058	,010
10	,394	,125	,394	,039	,048
11	,846	,019	,077	,039	,019
12	,087	,087	,548	,231	,048
13	,221	,154	,433	,144	,048
14	,192	,135	,385	,250	,039
15	,904	,067	,000	,010	,019
16	,058	,356	,029	,539	,019

Legend: for each question the correct item is underline in orange; the outside questions that are too difficult and not selective are underlined in blue and the high distracters are circled; the outside questions that are too easy and not selective are underlined in violet,

The analysis of distracter index blended with the analysis of the facility index and the difficulty index suggests the revision of the questions n, 1, 3, 8, 11, 13, 15, Particularly the following figure shows a practical application of the data results to the textual revision,

Figure n. 4 - Example of practical application of the item analysis to the textual analysis: question n. 8

The question n. 8 analyze the reading literacy skill to do direct inference, obtaining information from one or several implicit information from the text and/or from the personal encyclopedia (please see Fig, 1).
It's a multiple choice question; the correct item is D.

The question was:
8. The air transport is convenient:
A. when you bring few goods (15,4% of responses)
B. when you transport many goods (6,7% of responses)
C. when you travel long distances (61,5 % of responses; index of distracter: 0,615)
D. when speed is necessary (12,5% of responses)

The item analysis shows that the 3,9% of student didn't answer to the question, Only 12,5% of students answered to the correct item, The index of facility of the item is ,125 so that the item is too difficult, The index of selectivity is ,059 so that the item is not selective, The main distracter is the response C (the value of the distracter index is ,615),
From this assumption the suggestion is to re-design the question,

5. Conclusion

In general the analysis showed that reading comprehension of narrative texts was easier than understanding informative texts (media or text book). Students had difficulties on: doing direct inferences; global reading comprehension (understanding the topic); reconstructing parts of the text; logical organization and cohesion of the text (referent pronouns, syntactic links) and lexical aspects (morphological changes, figurative meaning). Particularly the distracter analysis showed that students were not able to integrate or to select the right information (explicit or implicit) from the text (interpretation) instead of the concurrent ones. The evidence was that student assumed linear strategies for reading comprehension.
The findings of this paper have significance for practicing teachers and test designer in the since that particular care should be taken while selecting items in a test to achieve an accurate measurement of students' behavior.
Item analyses should be utilized to improve already existing tests instead of developing new items to avoid wastage in time.
This study can be replicated in other subjects' areas to develop a good and useful item bank for practical utility.

6. References

Boothroyd, R. A., McMorris, R.F. & Pruzek, R.M. (1992). What do teachers know about measurement and how did they find out? Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA. *ERIC Document Reproduction Service*, No. 351 309.

Carroll, T., & Moody, L. (2006). Teacher-made tests, *Science Scope*, 66–67.

Freeman Frank S. (1962). *Theory and Practice of Psychological Testing*. New Delhi: Oxford & Ibh publishing.

Gronlund, N. E. (1998). *Assessment of student achievement*. 6th edition. Boston: Allyn and Bacon.

Haladyna, TM. (2004). *MC formats*. In: Haladyna TM (ed). Developing and validating multiple-choice test item. 3rd edn. Mahwah, New Jersey: Lawrence Erlbaum Associates; 67–96.

Herman, J. L., & Dorr-Bremme, D. W. (1984). *Teachers and testing: Implications from a national study*. Draft. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. ERIC Document Reproduction Service No. 244987.

Invalsi (2011). *Quadro di riferimento della prova di Italiano*, Sistema Nazionale di Valutazione, Retrieved September, 2, 2011, from http://www.invalsi.it/snv1011/documenti/QdR_Italiano.pdf.

Lange, A., Lehmann, I.J. & Mehrens, W.A. (1967). *Using item analysis to improve tests. Journal of Educational Measurement*, 4(2), 65-68; http://www.jstor.org/stable/1434299.

Palumbo, M,, & Garbarino, E. (2006). *Ricerca sociale: metodo e tecniche*. Milano: Franco Angeli.

Popham, W. J. (2008). *Classroom Assessment: What Teachers Need to Know* (5th ed.). Boston: Allyn and Bacon.

Remmers H.H., Gage, N.L. & Rummel, J.I. (1967). *A Practical introduction to measurement and evaluation* (2nd ed.). Delhi: Universal Book Stall.

Sireci, S. & Parker, P. (2006). *Validity on Trial: Psychometric and Legal Conceptualizations of Validity*. Article first published online: 15 NOV 2006; DOI: 10.1111/j.1745-3992.2006.00065.

Sharma, S.R. (2000). *Modern teaching strategies.* New Delhi: Omsons Publications.

Stiggins, R.J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice*, 10, 7-12.

Stiggins, R. J. (2001). *Student-involved classroom assessment* (3rd ed.). New Jersey: Prentice Hall.

Swanson, D.B., Holtzman K.Z., Albee K, & Clauser, B.E. (2006). Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Acad Med*, 81, 52–55.

Swanson, D.B., Holtzman K.Z., Clauser, B.E., Sawhill, A.J. (2005). Psychometric characteristics and response times for one-best-answer questions in relation to number and sources of options. *Acad Med*, 80, 93-96.

Trice, A. D. (2000). *A handbook of classroom assessment*. New York: Longman.

Walsh, K. (2008). *Advice on writing multiple choice questions (MCQs).* BMJ Careers 2005. Available at http://careers.bmj.com/careers/advice/view-article.html?id=616, accessed on 23 May 2008.

Walsh, K. (2008). *Answering multiple choice questions.* BMJ Careers 2005. Available at http://careers.bmj.com/careers/advice/view-article.html?id=891 (accessed on 23 May 2008).

Williams, J. M. (1991). Writing quality teacher-made tests: A handbook for teachers. *ERIC Document Reproduction Service*, No. 349 726.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. Journal of Teacher Education, 42, 37-42.

Xu, Y., & Liu, Y. (2009). Teacher assessment knowledge and practice: a narrative inquiry of a chinese college EFL teacher's experience. *TESOL Quarterly*, 43(3), 493–513.